Q1: Consider the following mini corpus containing 4 documents: [2 points]

Doc 1: new home sales top forecasts
Doc 2: home sales rise in july
Doc 3: increase in home sales in july
Doc 4: july new home sales rise

Compute the inverted index for this collection as discussed in class.

Q2: What is the time complexity \(O\) in terms of the length of the posting lists for the most efficient way of computing/retrieving all documents for the following queries? [4 points]

(a) \(a \text{ AND } b\)
(b) \(\text{NOT } a\)
(c) \(\text{NOT } b\)
(d) \(a \text{ OR } b\) (assuming you have a hash set data-structure which can perform lookups in \(O(1)\), i.e., constant time)

For terms/words, \(a\), \(b\), the length of the posting list is \(L(a), L(b)\). So your final answer should be in big-\(O\) notation. Something like \(O(g(\cdot))\): where \(g(\cdot)\) is a function of \(L(a), L(b)\). Also assume that the total number of documents in the corpus is \(N\). Show/give reasons as to how you arrived at your solution.

Q3. Recommend the most efficient query processing order for the following query: [2 point]

kaleidoscope AND tangerine AND marmalade AND trees

Assume that these terms in a given corpus have the following length of their posting lists:

<table>
<thead>
<tr>
<th>Term</th>
<th>Size of posting list</th>
</tr>
</thead>
<tbody>
<tr>
<td>kaleidoscope</td>
<td>87009</td>
</tr>
<tr>
<td>marmalade</td>
<td>107913</td>
</tr>
<tr>
<td>tangerine</td>
<td>46653</td>
</tr>
<tr>
<td>trees</td>
<td>316812</td>
</tr>
</tbody>
</table>

Q4: For what value of \(\rho\) and \(B\) does the Mandelbrot’s law \(f = P(r + \rho)^{-B}\) becomes the Zipf’s law? [2 points]